Hardware y Software de tres Supercomputadoras

Matías Zabaljáuregui

0. Introducción

El presente informe intenta describir, con cierto nivel de detalle, a tres de las supercomputadoras más potentes del mundo. Se seleccionaron de entre las primeras cinco posiciones del ranking que el sitio web www.top500.org publicó en noviembre del 2004 (ya existe un nuevo ranking, pero las máquinas elegidas siguen estando entre las primeras 5).

También se tuvo en cuenta, tal vez como factor excluyente, el sistema operativo que estas máquinas ejecutan. Para este informe, se decidió investigar el soporte que ofrecen los sistemas operativos y aplicaciones de código abierto/software libre, en particular, la versión 2.6 del kernel del sistema operativo gnu/linux, el cual incluye una serie de mejoras importantes relacionadas con la escalabilidad y está siendo utilizado por el 60 % de las supercomputadoras presentes en la lista de www.top500.org.

Dado que la descripción en detalle de todos los componentes de una máquina de este tipo puede llegar a tener una longitud considerable, se decidió hacer incapié en lo que se supone influye en mayor medida tanto al diseñador del sistema operativo, como al programador en cuanto al paradigma de parogramación a elegir: tipo y cantidad de procesadores, tipo y distribución de la memoria principal, y por último, tecnologías de interconexión. Por lo tanto, se ignoraron temas que, si bien son interesantes, no son relevantes para el presente estudio: consumo eléctrico, ventilación, infraestructura de almacenamiento en línea o fuera de línea, etc.

La descripción del hardware se completó con la mencion del sistema operativo y librerías que cada máquina utiliza. Y se intentó una conclusión de cuales son los modelos de programación paralela que mejor se adaptan a cada una de las arquitecturas mencionadas. En la segunda parte del informe, se intentará hacer una descripción de las modificaciones de diseño e implementación que se realizaron al kernel linux 2.6 para soportar las característica específicas de estas supercomputadoras.

La descripción para cada máquina intenta una cierta estructura común. Para comenzar se mencionan las características generales, y se listan sus componentes principales. Luego se estudian en detalle los microprocesadores, memoria y tecnologías de red. Finalmente se hace mención de los temas relacionados con el software que corre en la máquina.

El objetivo principal del informe es mostrar cuales son las tecnologías con las que se implementan las máquinas con mayor poder de cómputo del mundo, tanto en la capa de hardware como en la de software. Se intentó que los ejemplos correspondieran a distintos modelos teóricos de máquina paralela (SMP, ccNUMA, etc) y que tuvieran distintos modelos de programación (memoria compartida, pasaje de mensajes, etc). Sin embargo es evidente la tendencia hacia algún tipo de cluster como arquitectura y alguna variante de pasaje de mensajes como modelo de programación en la evolución de sistemas paralelos, por lo que las tres máquinas descriptas a continuación tienen muchas características en común, aunque no dejan de ser interesantes los detalles de implementación que diferencia a una de otras. Por lo tanto, en lugar de hacer extensa la lista de máquinas estudiadas (lo cual no agregaba diversidad al informe en términos de modelos teóricos) se prefirió sumar detalles a las descripciones.

1. BlueGene/L

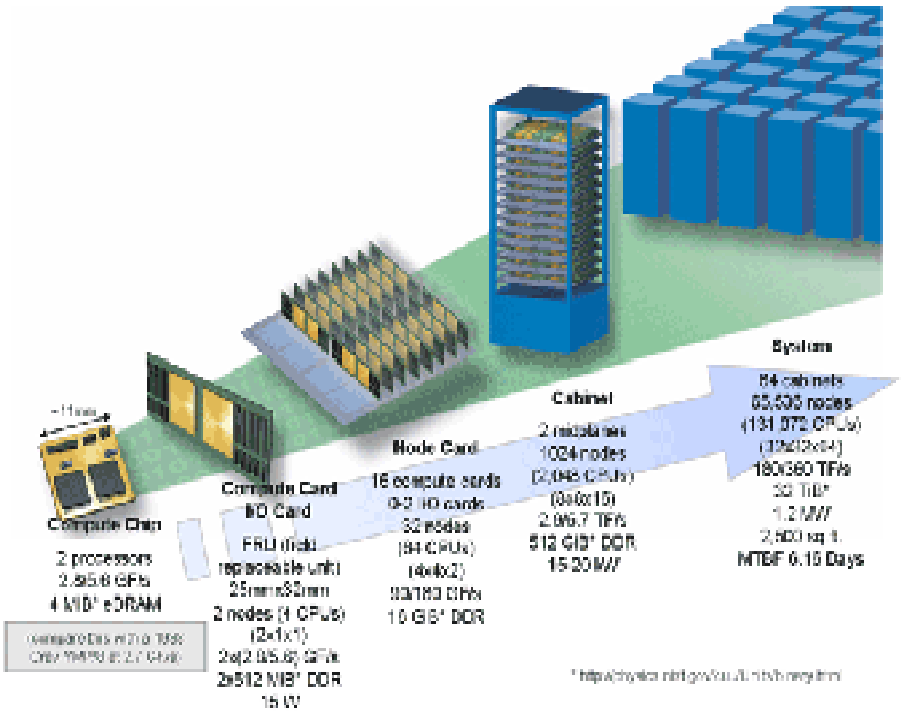
Descripción General

Blue Gene es un proyecto de investigación de arquitectura de computadoras diseñado para producir varias supercomputadoras de última generación operando en el rango de los PFLOPS [1]. Es un proyecto coperativo entre el departamento de defensa de los Estados Unidos, IBM y la academia. Hay cinco proyectos Blue Gene en desarrollo, entre ellos Blue Gene/L (tambien conocida como BG/L), Blue Gene/C, and Blue Gene/P, pero hasta ahora sólo el primero llegó a implementarse.

[1] Repasando las equivalencias: **megaFLOPS** (MFLOPS, 10^6 FLOPS), **gigaFLOPS** (GFLOPS, 10^9 FLOPS), **teraFLOPS** (TFLOPS, 10^{12} FLOPS), **petaFLOPS** (PFLOPS, 10^{15} FLOPS).

El sistema se construye a partir de un gran número de nodos, cada uno con una velocidad de reloj moderada (700 MHz) y bajo consumo eléctrico. Es un sistema escalable en el cual el máximo número de nodos asignados a una única tarea paralela es de 65.536 y cada nodo es un único ASIC (Application Specific Integrated Circuit o Circuito Integrado de Aplicación Específica) basado en la tecnología system-on-a-chip (SoC) de IBM que integra procesadores, memoria local (hasta 2 GB) y lógica de comunicación en un mismo chip.

El diseño de ensamblado de puede verse en en la siguiente figura.

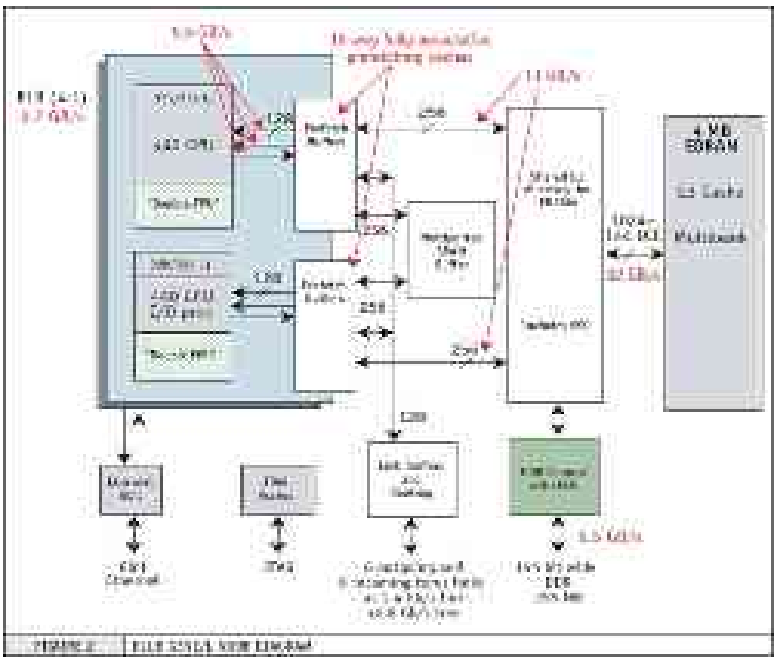


Como se puede observar, una plaqueta de cómputo contiene 2 nodos. Luego se ubican 16 plaquetas de cómputo en una "placa de nodo" y 1024 de estas placas en un gabinete. Por último, 64 gabinetes dan el total de 65.536 nodos de cómputo, es decir, 131.072 CPUs.

Además el sistema puede tener un número variable de nodos de entrada/salida (en este momento se dispone de un nodo E/S por cada 64 nodos de cómputo lo que da un total de 1024 nodos E/S), que se construyen con el mismo ASIC que los nodos de cómputo y tienen más memoria y conexiones gigabit Ethernet. Para compilar y analizar los resultados se requiere una computadora 'host'. Los nodos E/S manejan la comunicación entre los nodos de cómputo y otros sistemas, incluyendo la máquina host y los servidores de archivo.

Procesadores
El ASIC incluye dos núcleos de procesamiento PowerPC 440 de 32 bits, cada uno con dos FPUs (Floating-Point Units) de 64 bits. Cada núcleo tiene una cache de instrucciones y una cache de datos L1 de 32 KB cada una, una cache L2 de 2KB y una cache compartida de 4MB.

A una velocidad de reloj de 700MHz, la máxima performance teórica de un nodo es 2.8 Gflops usando sólo un núcleo y 5.6 Gflops cuando se usan ambos núcleos y las 4 FPUs. Esto le da al sistema entero una performance teórica de 180/360 TFLOPS.



El ASIC incluye dos núcleos de procesamiento PowerPC 440 de 32 bits, cada uno con un núcleo PowerPC 440 FP2 el cual es una unidad "doble" de punto flotante de 64 bits.
El 440 es un microprocesador superescalar estándar de 32 bits que suele ser usado en aplicaciones embebidas. Cada núcleo tiene una cache de instrucciones y una cache de datos L1 de 32 KB cada una, una cache L2 de 2KB y una cache compartida de 4MB.

Como no implementa el hardware necesario para proveer soporte SMP, los dos núcleos no son coherentes a nivel de cache L1 por lo que se utiliza un lock para permitir comunicación coherente entre procesadores.

En el modo de operación normal, un par CPU/FPU se usa para realizar computaciones mientras que el otro par es usado para el envio y recepción de mensajes.Sin embargo, no existen impedimentos de hardware para utilizar el segundo par CPU/FPU en algoritmos que tengan una baja relación entre computación/comunicación.

La unidad FP2 consiste en dos sub-unidades, una primaria y una secundaria, y cada una es escencialmente una unidad completa de punto flotante de 64 bits. Un conjunto ampliado de instrucciones incluye capacidades que superan a las arquitecturas tradicionales SIMD ya que una única instrucción puede iniciar operaciones, diferentes pero relacionadas, sobre datos distintos en cada una de las sub-unidades. A estas instrucciones se las denominan SIMOMD (por Single Instruction Multiple Operation Multiple Data).

Otra ventaja sobre las arquitecturas SIMD estandar es la habilidad de cada una de las sub-unidades de acceder a datos residentes en registros de la otra sub-unidad, lo cual logra notables mejoras de performance, especialmente en el tratamiento de números complejos. Además un camino de datos de 128

bits entre la cache de datos de la CPU y la FPU permite la transmisión simultánea de dos elementos de datos en un ciclo de reloj.

Red de Interconexión

Los nodos se interconectan a través de cinco redes:

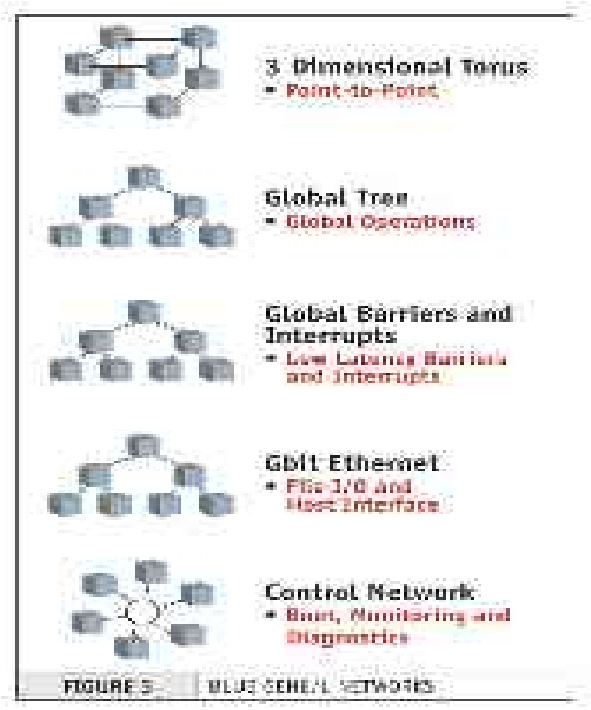
- una red torus 3D para pasaje de mensajes punto a punto y multicast (a una 'clase' seleccionada de nodos) de uso general. La topología se construye con links seriales entre los routers embebidos en los ASICs.

- un árbol global (combining/broadcast) para operaciones colectivas. El uso de registros programables flexibilizan el control de la red árbol. En su forma más simple, los paquetes que viajan hacia la raiz del árbol pueden ser punto a punto o combining. Los paquetes punto a punto se usan, por ejemplo, cuando un nodo de cómputo necesita comunicarse con su nodo E/S. Los paquetes combining se usan para soportar operaciones colectivas de MPI, como MPI_Allreduce, a través todos los nodos conectados al árbol (usando MPI_COMM_WORLD).

- una red de interrupciones y barreras globales sólo entre nodos de cómputo

- una red "Gigabit Ethernet to JTAG (Joint Test Action Group)" para control y monitoreo de la máquina, entre todos los nodos

- otra red Gigabit Ethernet para conexiones con otros sistemas como la máquina host y los servidores de archivos y sistemas front-end.



Por cuestiones de costo y eficiencia, los nodos de cómputo no están directamente conectados a la red gigabit ethernet, en cambio utilizan el global tree para comunicarse con sus nodos I/O, los cuales a su vez utilizan la red Gigabit Ethernet para comunicarse con los otros sistemas.

Ademas de los tipos de nodos ya vistos, existen los ASIC "link". Cuando las redes torus, arbol y de interrupciones cruzan el límite de la "placa de nodo", pasan a través de un ASIC especial, el cual tiene dos

funciones. Primero, regenera las señales sobre los cables entre placas de nodos, lo cual preserva la forma y amplitud de la señal. Segundo, el ASIC del tipo link puede redirigir las señales entre diferentes puertos, lo cual permite que BG/L pueda ser particionada en múltiples sistemas lógicos entre los cuales no habrá interferencia de tráfico. Cada partición formada de esta manera tendrá su propia red torus, arbol y de interrupciones asiladas unas de otras.

Software

Como se dijo anteriormente, los sistemas de código abierto estan siendo la elección preferida para las grandes máquinas paralelas. Este es el caso con BG/L, donde los nodos de I/O y front-end corren Linux y los nodos de cómputo ejecutan un kernel inspirado por el diseño de Linux. Otros componentes fundamentales son las implementaciones de MPI y MPICH2, que poseen una licencia BSD y son el modelo de programación principal de BG/L. La elección de utilizar Linux se basó principalmente en que la comunidad científica esta muy familiarizada con el sistema operativo, sus librerías e interfaces.

Sistema Operativo

Una aplicación para BG/L se organiza como una colección de procesos, cada uno ejecutándose en su propio nodo de cómputo de una de las particiones del sistema. Los nodos I/O proveén servicios adicionales. Se decidió dividir la funcionalidad del sistema operativo entre los nodos de cómputo y los nodos I/O. Cada nodo de cómputo se dedica a ejecutar un único proceso mientras que los nodos I/O proveén la interface física al sistema de archivos y permiten disparar los procesos y realizar operaciones de control.

Gracias a esta división, el kernel que se ejecuta en los nodos de cómputo es simple, liviano, y soporta la ejecución de una única aplicación con a lo sumo dos threads. El kernel se complementa con una librería en espacio de usuario que proveé a los procesos el acceso directo a las redes torus y árbol. Juntos, el kernel y la librería implentan la comunicación entre nodos de cómputo a través de la red torus y la comunicación entre nodos I/O y nodos de cómputo a través de la red árbol. La primera se utiliza para el intercambio de datos de las aplicaciones y la segunda permite a las aplicaciones acceder a los servicios ofrecidos sólo en los nodos I/O.

Como se mencionó, los nodos I/O corren el sistema operativo Linux, soportando la ejecución de múltiples procesos. El propósito de estos nodos es el de complementar a las particiones de nodos de cómputo con servicios tales como el acceso a un sistema de archivos, conexiones usando sockets a procesos en otros sistemas, etc.

El software de la BG/L también incluye un conjunto de servicios de control que se ejecutan en el sistema host. Muchos de estos servicios, incluyendo la inicialización del sistema, el particionado virtual de la máquina, mediciones de performance del sistema, no son visibles al usuario y se realizan a través de la red JTAG (la cual también es invisible para el usuario).

Programación de BG/L

Existen varias aproximaciones para programar BG/L. Se incluye el modelo de pasaje de mensajes usando MPI con C, C++ o FORTRAN. Tambien se ofrecen modelos de programación con un espacio de direccionamiento global como Co-Array FORTRAN (CAF) and Uni- fied Parallel C (UPC). Por último, algunas librerías matemáticas se están actualizando para aprovechar las ventajas de BG/L.

- MPI (Message Passing Interface): El modelo de programación principal para BG/L es el pasaje de mensajes utilizand una implementación de la librería MPICH2 la cual se mapea eficientemente sobre las redes torus y árbol. MPI es un modelo de programación muy conocido y representa una forma fácil de migrar código de aplicaciones existentes a BG/L.

- Co-Array FORTRAN (CAF) y Unified Parallel C (UPC): CAF y UPC son lenguajes explícitamente paralelos y de espacio de direccionamiento global que incorporan el modelo SPMD en Fortran 90 y C respectivamente. Por ejemplo, UPC agrega la instrucción “forall” al lenguaje C para distribuir un loop de

tipo for entre los nodos. Aqui puede notarse el contraste con el modelo de pasaje de mensajes donde el programador debe especificar que datos son enviados a que nodos.

- Librerías Matemáticas: Algunas librerías matemáticas también están siendo actualizadas para aprovechar las ventajas de la arquitectura de BG/L, como por ejemplo la posibilidad de utilizar el segundo nucleo de ejecución para aplicaciones de cómputo intensivas. Entre ellas se encuentran:

    - La "Engineering Scientific Subroutine Library (ESSL)", una implementación de IBM del "Linear Algebra Package (LAPACK)".
    - The Mathematical Acceleration Subsystem (MASS), una librería matemática optimizada.
    - Las librerías "Fast Fourier Transforms (FFT)" y "3D-FFT", que también están siendo optimizadas para aprovechar las FPUs de BG/L.

Ejecutando Programas

BG/L ejecuta un programa a la vez; la imagen del programa se carga y ejecuta en los miles de nodos de cómputo que se encuentren disponibles. Pero existen dos aproximaciones para ejecutar este programa:

- El modelo de coprocesador: Cada nodo corre una instancia del programa (65.536 instancias) con dos threads (una por nucleo de ejecución) y una memoria compartida de 256MB.
- El modelo de nodo virtual: Los dos núcleos de cada nodo de cómputo cargan y ejecutan separadamente una imagen (sin multithreading) de un programa (131.072 instancias) utilizando 128MB cada una.

BG/L no es una máquina de propósito general. Su diseño está optimizado para resolver problemas basados en grid y que tienen como principales características una comunicación intensa entre vecinos cercanos y un requerimiento alto de poder de cómputo. En cambio, GB/L no se comporta de manera óptima en aquellos casos donde la comunicación entre nodos es escasa o nula.

2. COLUMBIA

Columbia, nombrada en honor a los astronautas del famoso transbordador, es la supercomputadora más poderosa de la NASA y da soporte a las investigaciones científicas realizadas en las misiones de la agencia. Esta supercomputadora se construye como un cluster de 20 máquinas SGI Altix 3700, cada una con 512 procesadores y una memoria compartida de 1 terabyte, lo que ofrece un total de 10,240 procesadores y una memoria principal de 20 terabytes. La capacidad de almacenamiento en línea es de 440 terabytes y el almacenamiento fuera de línea alcanza los 10 petabytes.

SGI Altix 3000 (Modelos Altix 3300, Altix 3700)

La familia SGI Altix 3000 combina la arquitectura NUMAflex con componentes estándares de la industria, como el procesador Itanium 2 de Intel y el sistema operativo GNU/Linux. La arquitectura NUMAflex permite al sistema escalar hasta los 512 procesadores con una aproximación ccNUMA, utilizando el concepto de memoria compartida global incluyendo procesadores y E/S de todos los nodos. Ésto evita el overhead de la comunicación por red o E/S y se logra en gran medida gracias a un sofisticado sistema de interconexión de memoria denominado NUMAlink.

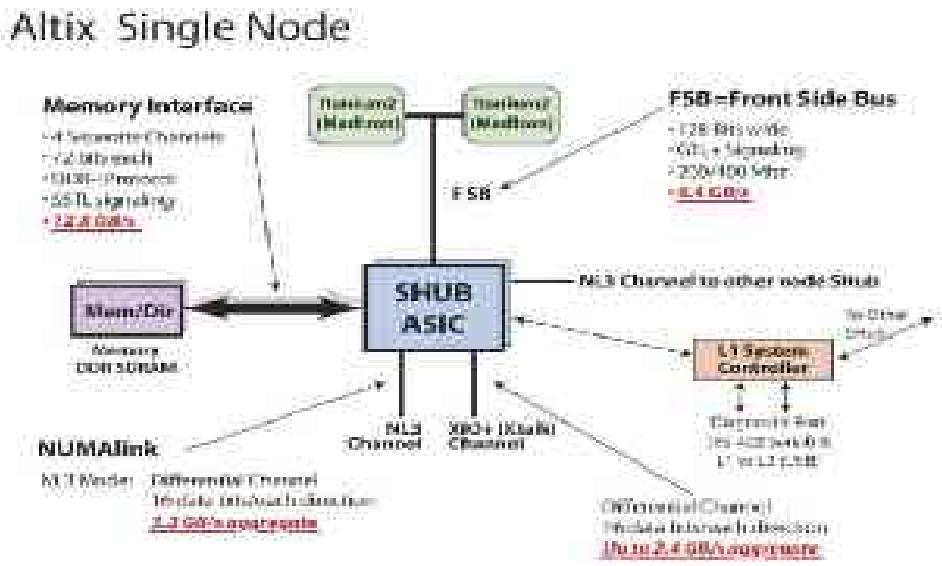
Arquitectura NUMAflex

Este diseño permite que la CPU, memoria, E/S, interconexión, dispositivos gráficos y almacenamiento sean distribuidos en componentes modulares o “bricks”. Además utiliza un protocolo llamado SGI NUMA implementado directamente en hardware para permitir una gran performance y el diseño modular.

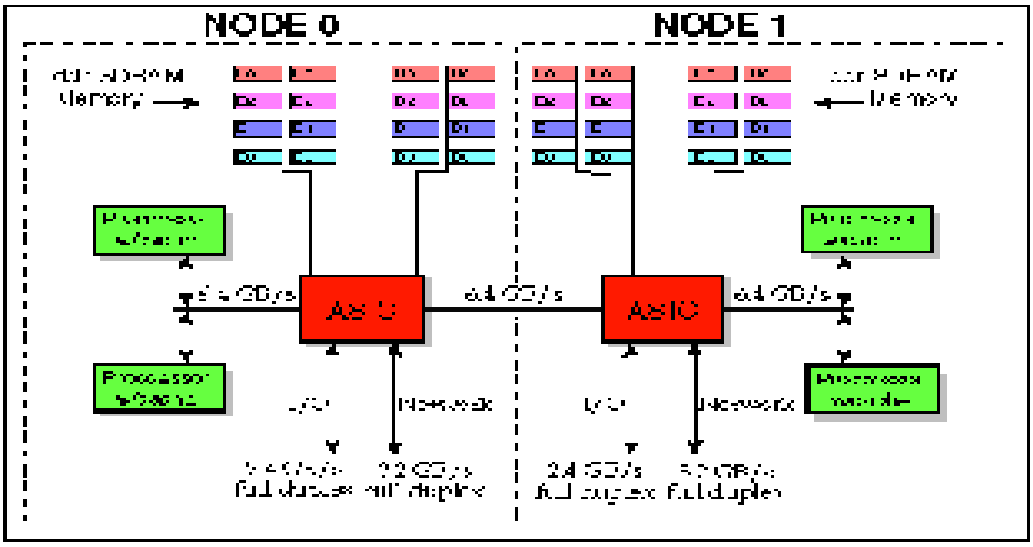
Cada C-Brick de Altix 3000 contiene:

- Nodos y Procesadores: Dos nodos en cada C-Brick, con dos procesadores Itanium 2 (Madison) en cada nodo para un total de 4 procesadores. Cada procesador contiene 32 KB de cache L1, 256 KB de L2 y 6 MB de cache L3.
- SHUB: Dos ASICs denominados “Hubs Escalables” (Scalable Hub o SHUB); cada SHUB se conecta con 2 procesadores de un nodo, memoria, dispositivos E/S y otros SHUBs. El protocolo de coherencia de cache de la Altix se implementa en el SHUB, esto se explica más abajo. Dentro del C-Brick, los dos SHUBs se conectan por un canal de 6.4 BG/segundo. Un puerto externo de 3.2/6.4 GB/segundo en cada SHUB se conecta con un SHUB en otro C-Brick o a alguno de los "planos de routers". Los planos de routers se construyen con routers de 8 puertos estructurados en una topología de tipo fat-tree.
- Memoria Principal Local: Hasta 32 GB de memoria por C-Brick (En el sistema de la NASA, cada C-Brick tiene aproximadamente 7.6 GB de memoria)
- Entrada/Salida: Se incluye una interface de red y una de Entrada/Salida. Se usan para conectar los cables de la interconexión NUMALink, que a la ves integra a cada C-Brick en la red de interconexión y para conectar a un Brick de E/S (ver más abajo).

A continuación se muestra el diagrama de un nodo en detalle:



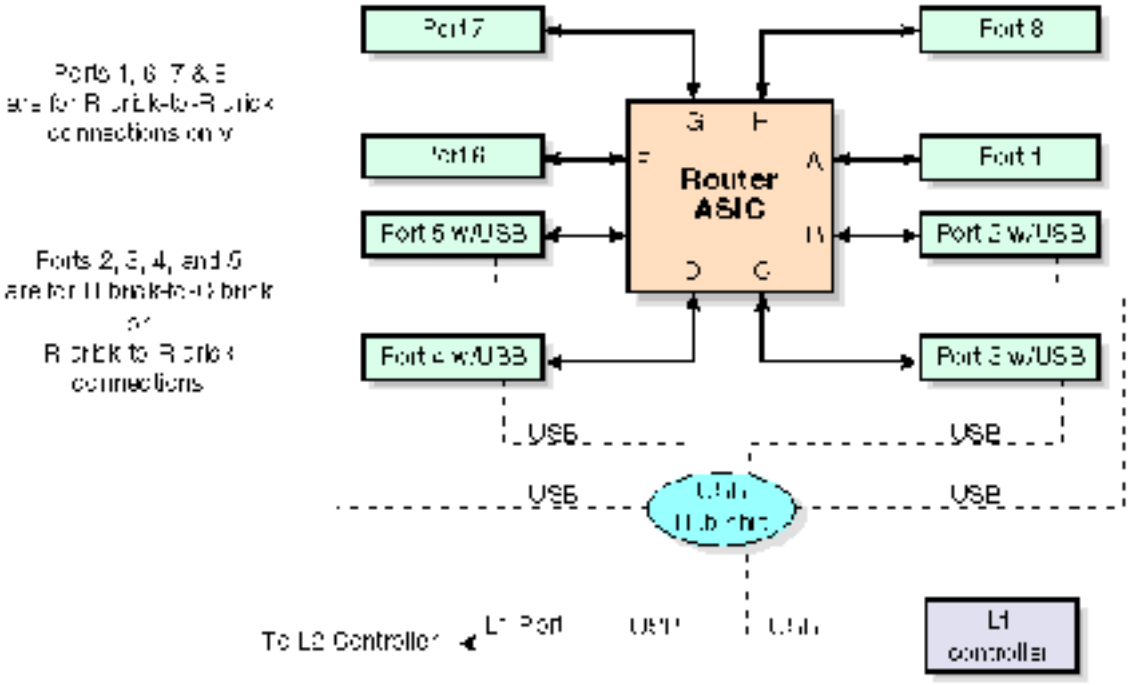
Como se dijo anteriormente, un C-Brick se construye a partir de dos nodos:



El resto de los componentes de un sistema Altix son:

- El R-Brick, un router NUMAlink 3 de ocho puertos, que se usa para construir la red de interconexión entre los C-Bricks.
- El M-Brick, es un módulo de memoria que permite escalar la cantidad de memoria principal indpendientemente del resto de los Bricks de la red de interconexión.
- El IX-Brick, es el módulo de E/S básica y el PX-Brick es un módulo de expansión PCI-X que se conecta al C-Brick por medio del canal E/S.

Un R-Brick en detalle:



Los 20 sistemas Altix 3000 de la serie 3700 que conforman el COLUMBIA se denominan Columbia1 a Columbia20 y a estos se les suma una máquina extra (denominada sencillamente "Columbia") como front end de la supercomputadora.

Columbia (sistema front end):

- 16 C-Bricks
- 8 R-Bricks
- 1 Ix-bricks

Columbia1 – Columbia12:

- 128 C-Bricks
- 112 R-Bricks
- 3 Ix-bricks

Columbia13 - Columbia20:

- 64 C-Bricks
- 48 R-Bricks
- 4 Ix-bricks

Procesadores

El chip Itanium está basado en la arquitectura IA-64 (Intel Architecture, 64 bit) la cual implementa la tecnología EPIC (Explicit Parallel Instruction set Computing). Con EPIC, un compilador para la familia Itanium convierte el código secuencial en grupos de 128 bits de instrucciones que pueden ser directamente procesadas por la CPU sin que se necesite una interpretación adicional. Esta expresión explícita de paralelismo permite que el procesador se concentre en ejecutar el código paralelo lo más rápido posible, sin tener que preocuparse por cuestiones de optimización o interpretación extras. En cambio, los compiladores regulares (los que no generan código compatible con EPIC) toman código secuencial, lo examinan y

optimizan para ser paralelizado, pero luego deben regenerar código secuencial de forma tal que el procesador, en el momento de la ejecución, pueda re-extraer la noción de paralelismo del código generado. Por lo tanto el procesador tiene que leer este paralelismo implícito del código máquina, reconstruirlo y ejecutarlo.

El procesador Itanium usa instrucciones con una longitud considerable. Específicamente, como ya se mencionó, se agrupan tres instrucciones de 41 bits en grupos paralelizados de 128 bits. Los 5 bits más significativos del grupo codifican metainformación que incluye las unidades de ejecución que se usarán (unidades de enteros, unidades de memoria, unidades de punto flotante y unidades de bifurcación).

El Itanium2 puede ejecutar dos grupos paralelizados por ciclo. Pueden incluirse cuatro operaciones de carga desde la cache L2 a los registros de las dos unidades de punto flotante, por lo que se soportarán dos operaciones de punto flotante en un ciclo.

Es conocida la capacidad de los procesadores modernos de hacer predicciones en las bifurcaciones. En lugar de esperar a tener el resultado para saber qué camino de código seguir ejecutando, el procesador intenta predecir el resultado y continúa ejecutando una de las ramas. Si la prediccion resultara ser correcta, la ejecución ya realizada se valida, en cambio si la predicción fallara, debe comenzarse desde el inicio de la otra rama descartando todos los cálculos realizados. Sin la predicción de bifurcaciones, el paralelismo sería prácticamente imposible, por eso la familia Itanium minimiza el tiempo de ejecución siguiendo ambos caminos alternativos, y cuando se termina de calcular el resultado de la condición se desecha la rama inválida y se continúa con la rama válida.

Red de Interconeción

Columbia cuenta con las siguientes tecnologías de red para interconectar sus nodos:

- **SGI® NUMAlink™**
- **InfiniBand network**
- **10 gigabit Ethernet**
- **1 gigabit Ethernet**

A continuación se describen las dos primeras. Se decidió no describir al protocolo Ethernet por ser este un protocolo de capa de enlace (capa 2 del modelo osi) muy popular (y por lo tanto, muy documentado) y por no ser una tecnología de interconexión originalmente pensada para procesamiento paralelo.

SGI® NUMAlink™

Las Altix 3000 usan una nueva tecnología de comunicación denominada “canales NUMAlink 4” que alcanza un alto rendimiento utilizando técnicas avanzadas de señalización bidirecional. La red se configura en una topología “fat-tree” la cual permite que la performance del sistema escale bien proveyendo un incremento lineal del ancho de banda biseccional cuando el sistema crece en tamaño. Las Altix 3000 también ofrecen redes "dual-plane" o paralelas.

La siguiente figura muestra esta topología para una configuración de 512 procesadores. Los círculos en la figura representan R-bricks, las líneas representan cables NUMAlink y los 128 pequeños cuadrados a lo largo del centro del diagrama representan los C-bricks.

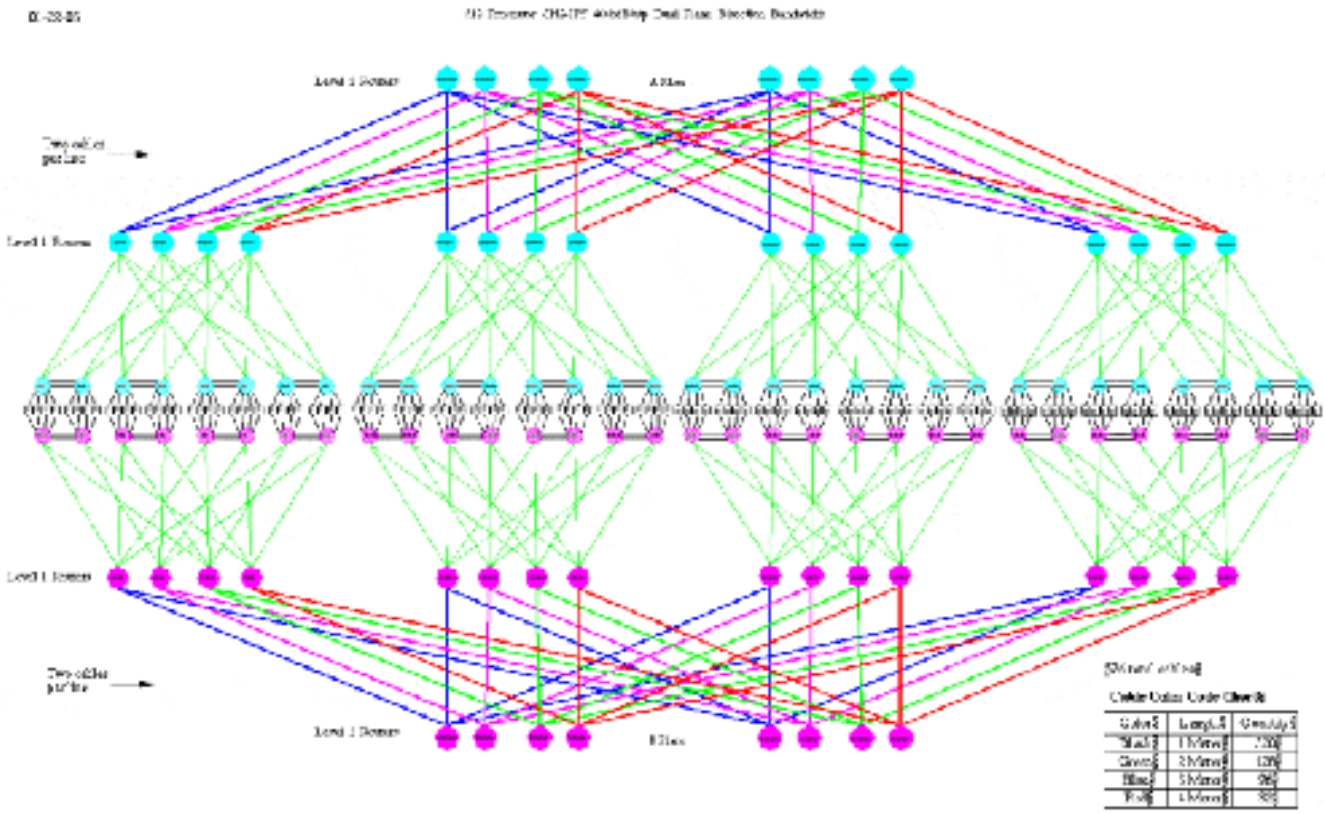


Figure 2. 512-Processor Dual "Fat-Tree" Interconnect Topology

InfiniBand

InfiniBand, conocido inicialmente como System I/O, nació en 1999 de la fusión de los proyectos Future I/O (desarrollado por Compaq, IBM y Hewlett Packard) y Next Generation I/O (desarrollado por Dell, Hitachi, Intel, NEC, Siemens y Sun Microsystems). En un principio, Infiniband se perfiló como un posible sustituto del bus PCI. Sin embargo, Infiniband está optimizado para ser utilizado en redes, y es innecesariamente complejo para una tecnología de entrada/salida local.

La arquitectura de InfiniBand es un estándar que define un nuevo subsistema de interconexión a alta velocidad punto a punto basado en switches. Este nuevo sistema de interconexión deja atrás el modelo de entrada/salida basada en transacciones locales a través de buses para implantar un nuevo modelo basado en el paso remoto de mensajes a través de canales. Esta arquitectura es independiente del sistema operativo y del procesador de la mquina.

El ancho de banda básico de un enlace simple (1x) es de 2.5 Gbits/s (320 MB/s). Cada enlace simple puede tener 1, 4 ó 12 líneas de conexión, consiguiéndose velocidades de transferencia bidireccionales (permite una comunicación full duplex entre dispositivos) de 5 Gbits/s (1x), 20 Gbits/s (4x) y 60 Gbits/s (12x), respectivamente . Infiniband permite transmitir datos entre sistemas a grandes distancias, que van desde 17 metros (conexiones con cableado de cobre) hasta cerca de 10 km (en conexiones con fibra óptica).

Las redes Infiniband se basan en switches que interconectan los distintos dispositivos. De esta forma, tenemos dos tipos de nodos: los intermedios (switches) y los finales (los distintos dispositivos de E/S, etc). Cada switch (o nodo intermedio) gestiona la comunicación entre los distintos subsistemas que interconecta, permitiendo múltiples envíos de datos simultáneos. Existe otro tipo de nodo intermedio (nodo router) que permite la interconexión de varias subredes Infiniband entre sí. Cuando se habla de subred se refiere a sistemas completos, unidos entre sí mediante este nuevo tipo de nodo. De este modo, la interconexión

Infiniband alcanza un nuevo nivel, permitiendo la interconexión de sistemas completos a largas distancias.

Software

El sistema operativo que corren las máquinas es un RedHat 7.2 con un kernel para procesadores Itanium, con algunas modificaciones para facilitar una única imagen del sistema en 64 CPUs y para incrementar la performance de entrada/salida en estas máquinas. También se utilizan un conjunto de librerías para computación paralela altamente optimizadas para explotar las herramientas y drivers de SGI para hardware NUMA.

Los diseñadores de SGI han escalado el sistema operativo Linux hasta el nuevo límite de 64 procesadores para una única imagen del kernel en una Altix 3000. En un cluster tradicional, los diseñadores deben proveer a cada nodo con la máxima cantidad de memoria que un proceso pueda necesitar y debe sumarse la memoria requerida por el sistema operativo en cada host. En cambio, con un tamaño de host mayor se dispone de una mayor cantidad de memoria para cada proceso individual corriendo en ese host. Por ejemplo, un kernel Linux corriendo en 64 procesadores tendrá 4 Tbytes de memoria disponible para un único proceso.

La arquitectura NUMAflex soporta la capacidad de tener múltiples nodos en una única red NUMAlink. Cada nodo será un sistema independiente corriendo su propia imagen del kernel Linux, por lo tanto si un nodo sufre un error fatal, los otros pueden continuar operando mientras el sistema reinicia el nodo que falló. Es posible utilizar firewalls para permitir o denegar accesos a memoria, CPU o I/O que crucen los límites entre nodos. También se dispone de componentes llamados BTEs (Block Transfer Engines) que operan como unidades DMA con coherencia de cache, y se utilizan para copiar datos de un rango de memoria física a otro a un ancho de banda muy elevado, incluso cruzando los límites entre nodos.

El acceso a memoria entre nodos (internode shared-memory o XPMEM) permite que un proceso pueda acceder a la memoria de otro proceso en una misma Altix 3000, sin importar si esta memoria reside en el mismo nodo o en uno separado. Esto puede realizarse utilizando BTE o directamente compartiendo la memoria física subyacente. La figura siguiente muestra la pila de protocolos que permitió a los diseñadores construir las capas de software para memoria compartida (XPMEM) y red (XPNET).

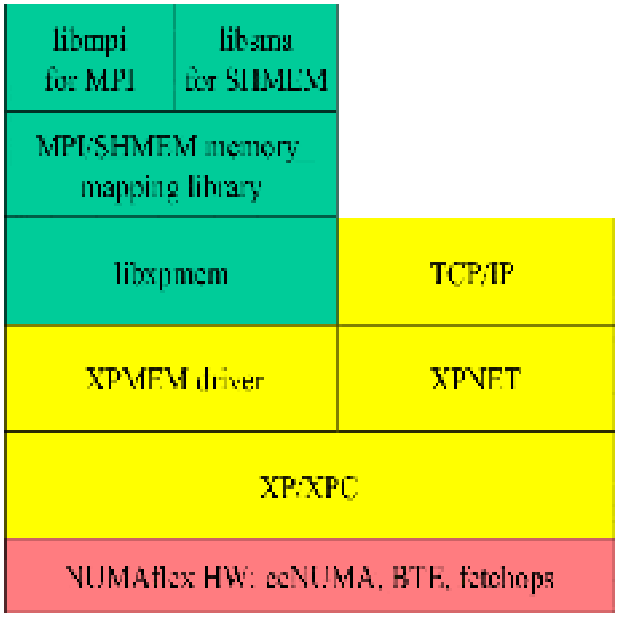


Figure 5. Software Stacks for XPMEM and XPNET

Los módulos del kernel XP y XPC proveén un canal de comunicación entre nodos que transmite datos sobre

la red NUMAlink. XPNET utiliza NUMAlink para proveer TCP y UDP al usuario. La librería de usuario libxpmem y el módulo del kernel XPMEM proveén acceso a memoria entre nodos.

El toolkit de pasaje de mensajes de SGI (MPI + SHMEM) está optimizado para usar XPMEM y en el futuro se preveé incluir interfaces similares a la interfaz System V (ampliamente utilizada en sistemas Unix) de memoria compartida. A continuación se mencionan algunas de las librerías disponibles que fueron altamente optimizadas para esta plataforma:

- Librerías Matemáticas:
    - libimf.a: Librería matemática de Intel, para ciertas aplicaciones específicas.
    - libm.a: Librería matemática de propósito general compatible con gnu.
    - SCSL: Funciones matemáticas y científicas optimizadas.

- MPT: El Message Passing Toolkit de SGI es un conjunto optimizado de librerías de programación paralela que incluye a MPI y SHMEM. Se incluyen las siguientes características relacionadas con MPI:
    - Compatible con MPI 1.2
    - Optimizadas las operaciones colectivas de MPI para NUMAlink
    - Implementadas algunas capacidades de MPI-2

- Con respecto a la librería SHMEM, se puede mencionar:
    - Son interfaces de memoria compartida sencillas
    - Implementación optimizadas
    - Capacidad de punteros globales, incluso entre nodos

3. MareNostrum

El Ministerio de Educación y Ciencia, la Generalitat de Catalunya y la Universidad Politécnica de Catalunya, tomaron la iniciativa de crear el Centro Nacional de Supercomputación, el hogar de MareNostrum, a principios del 2004. La misión del centro es investigar, desarrollar y gestionar las Tecnologías de la Información, a fin de facilitar el progreso científico. Con el objetivo de cumplir esta misión, se presta una especial atención a áreas como Supercomputación y Arquitectura de Computadores entre otras. En Marzo de 2004, el gobierno español y la empresa IBM, firmaron un acuerdo para construir el ordenador más rápido de Europa y uno de los más rápidos del mundo conocido actualmente como MareNostrum y administrado por el centro.

Descripción General

MareNostrum es un supercomputador basado en procesadores PowerPC, la arquitectura BladeCenter de IBM, el sistema operativo abierto Linux, y la red de interconexión Myrinet. Estas cuatro tecnologías configuran la base de una arquitectura y diseño que tendrán un gran impacto en el futuro de la supercomputación.

El resumen del sistema es el siguiente:

- 42.35 Teraflops de rendimiento de pico teórico.
- 4.812 procesadores PowerPC 970FX en 2406 Nodos duales
- 9.6 TB de memoria
- 236 TB de almacenamiento en disco
- 3 redes de interconexión
    - Myrinet
    - Gigabit Ethernet
    - Ethernet 10/100

MareNostrum está formado por 40 bastidores (racks) y ocupa 120 m^2. A continuación se mencionan los diferentes tipos de bastidores actualmente instalados.

Bastidores de Proceso

MareNostrum dispone de 29 bastidores dedicados a proceso, es decir, dedicados a cálculo. Estos 29 bastidores contienen un total de 4812 procesadores PowerPC 970FX trabajando a una frecuencia de 2,2 GHz y un total de 9.6 TB de memória. Cada uno de los bastidores de proceso está formado por 6 Blade Centers (explicado más adelante). En total, cada bastidor dispone de un total de 168 procesadores y 336 Gb de memória principal. Cada uno de los bastidores tiene aproximadamente una potencia pico teórica aproximada de 1,4 Tflops.

Un Blade Center esta formado por 14 nodos duales (conocidos como JS20) con un total de 28 procesadores y dispone de dos fuentes de alimentación redundantes de esta manera si falla una de las fuentes de alimentación el sistema puede seguir funcionando. También dispone de un switch para la red Gigabit que interconecta a los diferentes nodos JS20.

Por otro lado, cada uno de los nodos dispone de una tarjeta de red Myrinet tipo M3S-PCIXD-2-I para su conexión a la red de interconexión de alta velocidad. A pesar de que cada uno de los nodos dispone de un disco local de 40 Gb, cada nodo funciona de forma *diskless*, es decir, el sistema operativo no reside en el disco local, sino que reside en los bastidores de almacenamiento y en tiempo de inicialización del nodo, se carga en cada uno de ellos a través de la red Gigabit.

Bastidores de Red Myrinet

Los 2406 nodos JS20 y los 40 nodos que actúan como servidores de disco (nodos p615 en los bastidores de almacenamiento) se interconectan a través de la red de interconexión de alta velocidad Myrinet, y utilizan cables de fibra óptica como medio físico.

De los 40 bastidores existentes en MareNostrum, cuatro de ellos se dedican a los elementos de red que permiten interconectar los diferentes nodos conectados a la red Myrinet. Estos cuatro bastidores se encuentran en el centro de la sala y reciben un cable de fibra óptica por cada nodo (de proceso o almacenamiento). Los elementos de red interconectan los diferentes cables permitiendo la interconexión punto a punto de los diferentes nodos.
Toda la red de Myrinet se interconecta mediante (detallado más abajo):

- 10 switches tipo Clos256+256
- 2 switches tipo Spine 1280

Bastidores de Almacenamiento
Además de el disco local de cada nodo con una capacidad de 40 GB, MareNostrum dispone de 20 nodos de almacenamiento distribuidos en 7 bastidores. Los 7 bastidores disponen actualmente, de un total de 560 discos de 250 GB lo que dan una capacidad total de almacenamiento externo de 140 TB. Estos discos están trabajando con GPFS (Global Parallel File System), que además de ofrecer una visión global del sistema de ficheros, permite el acceso paralelo al mismo.
Los 2406 nodos acceden a los bastidores de disco a través de la red de interconexión Myrinet lo que proporciona un acceso a disco de alta velocidad.

Cada uno de los 20 nodos de almacenamiento dispone de dos nodos p615 responsables de servir las peticiones a disco, una controladora tipo FAStT100 y una unidad de expansión EXP100. Cada uno de los nodos p615 dispone de una conexión a la red Myrinet, una conexión a la red Gbit y una conexión a la red ethernet 10/100.

Bastidor de Operaciones
Uno de los bastidores es el bastidor de operaciones desde donde se gestiona el sistema. En este bastidor se encuentra la consola de la máquina. El contenido del bastidor es el siguiente:

- 1 Monitor 7316-TF3
- 2 nodos de gestión p615
- 2 HMC (consolas) 7315-CR2
- 3 Nodos remotos asíncronos
- 1 Chasis BCIO BladeCenter IO
- 4 Switches Cisco 3550

Bastidor de Comunicaciones
Uno de los bastidores de MareNostrum se dedica a los elementos de interconexión de las redes Gigabit y una parte de los elementos de interconexión de la red Ethernet 10/100 (el resto de elementos se encuentra en el bastidor de operaciones). Los elementos son:

- 1 Switch Force10 E600 Gigabit Ethernet**:**
  Dispone de 7 slots de los que 6 están ocupados con tarjetas10/100/1000 Base-T de 48 puertos ofreciendo un total 288 puertos GigabitEthernet / IEEE 802.3
- 4 Switches Cisco 3550 48-port Fast Ethernet:

Tecnologías BladeCenter, el PowerPC 970FX y la red Myrinet en detalle
Como se dijo anteriormente, los nodos de cálculo son 2.406 servidores BladeCenter JS20 de IBM ubicados en 163 chasis BladeCenter. Cada servidor tiene dos procesadores PowerPC 970 corriendo a 2.20GHz. La tecnología BladeCenter ofrece la mayor densidad comercialmente disponible (según IBM), lo que resulta en una alta performance con un menor espacio físico requerido. Se soportan hasta 84 procesadores duales en

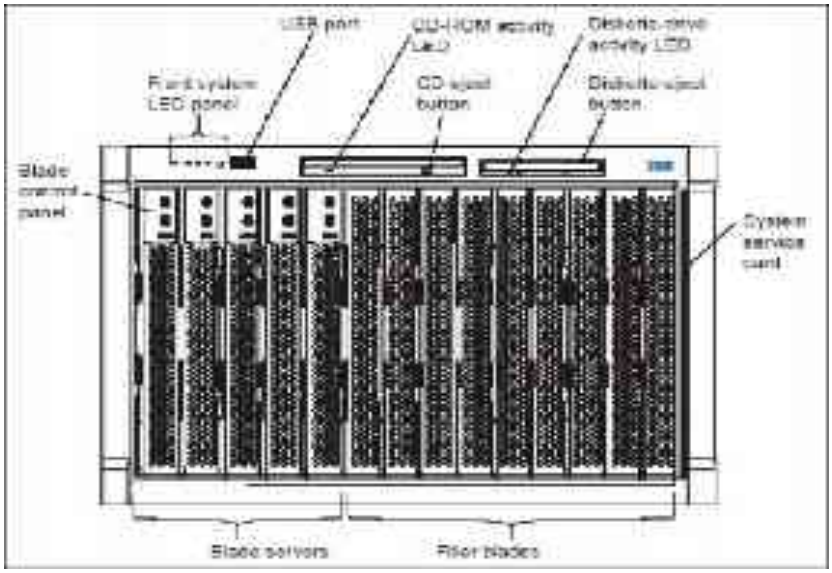
un único rack 42 U, por lo que cada rack agrega 1.4 teraflops de poder de cómputo.

El componente clave de la infraestructura BladeCenter es el chasis que puede contener varios dispositivos de conexión en caliente (hot-swappable) llamados *blades*. Los blades se ofrecen en dos variedades: *blades servidores* y *blades opcionales.*

Un *blade servidor* es un servidor independiente que contiene uno o más procesadores y memoria asociada, almacenamiento en disco, y controladores de red. Corre su propio sistema operativo y aplicaciones. Cada *servidor* se coloca en una bahia del chasis y se conecta a un backplane para compartir recursos comunes entre los que se encuentran la alimentación eléctrica, ventiladores, lectora de cdrom y diskettes, switches ethernet y Fibre Channel y puertos del sistema.

Los *blades opcionales* pueden ser compartidos por los *blades servidores* y proveén características adicionales, tales como controladores de entrada salida para arreglos de discos externos, alimentación eléctrica extra, etc.

A continuación se muestran una vista frontal y de atrás del chasis y se listan los módulos que pueden agregarse.

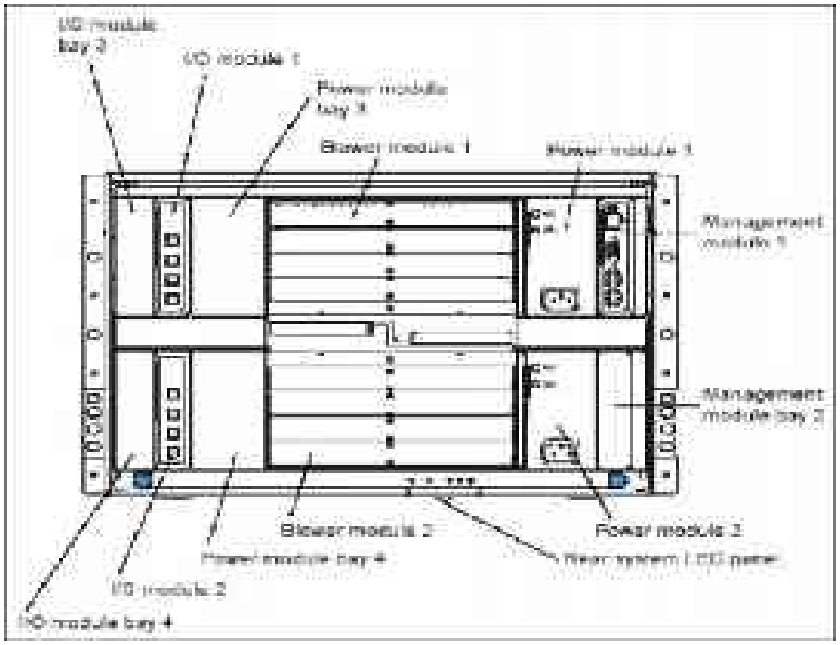


En la vista frontal, se pueden observar:

- Catorce bahías donde se pueden instalar los servidores.
- Una lectora de cdrom, una diskettera y un puerto USB que puede asignarse dinámicamente a cualquiera de los server blade instalados en el chasis.

La siguiente figura ilustra la parte trasera del chasis, donde pueden observarse los siguientes elementos:

- Un módulo de administración.
- Una bahía donde se puede instalar un módulo de administración redundante opcional.
- Cuatro bahías donde se pueden instalar módulos de entrada salida opcionales.
- Un par de módulos que proveén alimentación redundante.
- Dos bahías donde se pueden instalar módulos de alimentación redundante.
- Dos ventiladores.



Los módulos de entrada salida permiten la conectividad entre los servidores dentro del mismo chasis y entre los mismos servidores y el exterior. Los módulos disponibles proveén tres tipos de interconexión:

- Redes de área local (LAN)
- Redes de almacenamiento (SANs)
- Redes de alto ancho de banda y baja latencia usadas en clusteres (Myrinet)

La mayoría de los módulos soportan conectividad a un solo tipo de red, a excepción del módulo*pass-through*, el cual puede ser usado con los tres tipos de red. Cada módulo se conecta a cada servidor a través del backplane del chasis y para esto cada servidor debe tener una interface compatible. Actualmente disponen de interfaces gigabit ethernet que están conectadas a las bahías 1 y 2 del chasis. Por lo tanto en estas bahías sólo se podrá utilizar un módulo que implemente un switch de LAN o un módulo Pass-Thru.

Posibles modulos I/O:

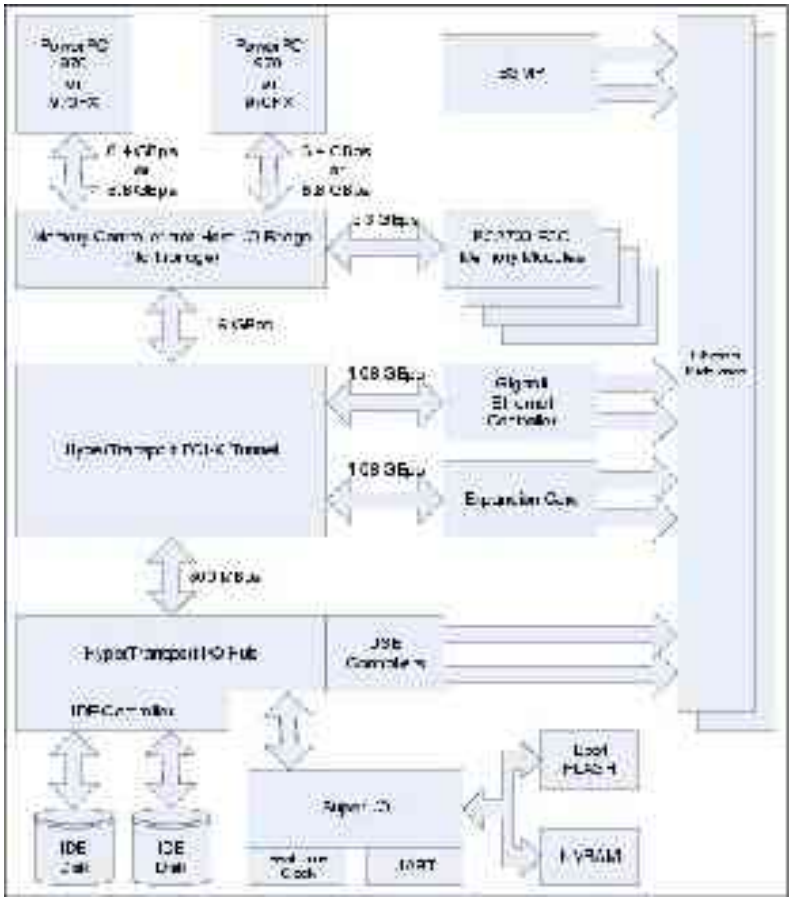
- IBM 4-Port Gigabit Ethernet Switch Module
- CISCO Systems Intelligent Gigabit Ethernet Switch Module
- Nortel Networks Layer 2-7 Gigabit Ethernet Switch Module
- 2-Port Fibre Channel Switch Module
- Brocade Enterprise SAN Switch Module
- Optical Pass-thru Module

A continuación se describe la arquitectura del server **BladeCenter JS20**, con el cual se construyen cada uno de los nodos de la supercomputadora MareNostrum.

El **BladeCenter JS20** es un *servidor blade* que ofrece las siguientes características:

- Dos microprocesadores PowerPC 970 o PowerPC 970FX a 1.6 GHz o 2.2 GHz, respectivamente.
- Hasta 4 GB de RAM
- Dos discos IDE
- Dos interfaces gigabit ethernet integradas
- Conectores de expansión para agregar las siguientes placas opcionales:
    - Dos interfaces gigabit ethernet adicionales
    - Dos interfaces Fibre Channel de 2 Gbps
    - Una interface Myrinet
- Un procesador de administración (Blade Systems Management Processor o BSMP)

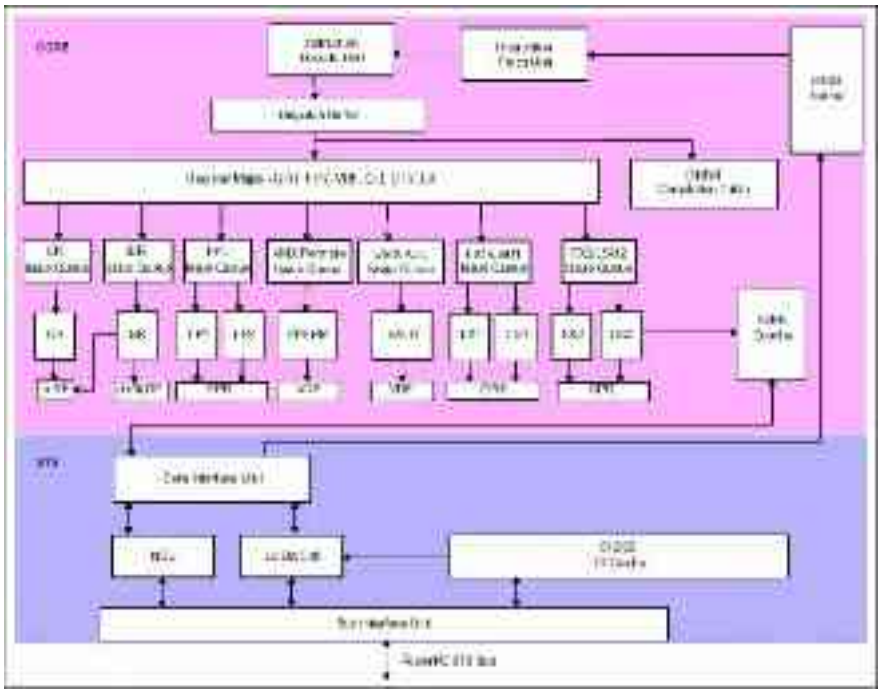
La siguiente figura ilustra la arquitectura del server:



Procesador PowerPC 970FX

El procesador PowerPC 970FX es un procesador con una arquitectura de 64 bits orientado propósito general (el Mac G5 de Apple incorpora este tipo de procesador). Se trata de un procesador superscalar con

extensiones vectoriales tipo SIMD (VMX) el diseño del cual se ha basado en el procesador de alta prestaciones Power4. A diferencia de su predecesor, el PowerPC 970 a 1.5 Ghz, el nuevo PowerPC 970FX basado en una tecnología de 90nm trabaja a una frecuencia de 2.2 GHz. Es capaz de lanzar hasta un máximo de 4 instrucciones por ciclo y es capaz de mantener hasta un máximo de 200 instrucciones en ejecución.



VMX

Vector/SIMD Multimedia eXtension (VMX) es una extensión a la arquitectura PowerPC. Define registros e instrucciones adicionales para soportar operaciones "single-instruction multiple-data (SIMD)" que aceleran la ejecución de tareas de cálculo intensivo. Se agrega un procesador vectorial llamado VXU al modelo de proceso lógico de PowerPC, el cual opera en vectores de 128 bits de longitud. Estos pueden ser interpretados por el VXU como:

- Un vector de 16 elementos de 8 bits
- Un vector de 8 elementos de 16 bits
- Un vector de 4 elementos de 32 bits

Estructura de la cache

El PowerPC 970 incluye varias caches on-chip para reducir la latencia cuando se buscan instrucciones o se realizan operaciones de cargas o almacenamiento de datos. A continuación se presenta una lista de ellas:

- Cache de instrucciones de 64 KB
- Cache de datos de 32 KB
- SLB (segment lookaside buffer) de 64 entradas
- TLB (translation lookaside buffer) de 1024 entradas

- Cache L2 de 512 KB

Características

El diseño de este procesador utiliza una variedad de técnicas para permitir una operación superescalar, donde múltiples instrucciones se ejecutan durante cada ciclo del reloj del procesador. Esta capacidad está soportada por el uso de múltiples unidades de ejecución en el núcleo del procesador.

Entre otras características se pueden mencionar:

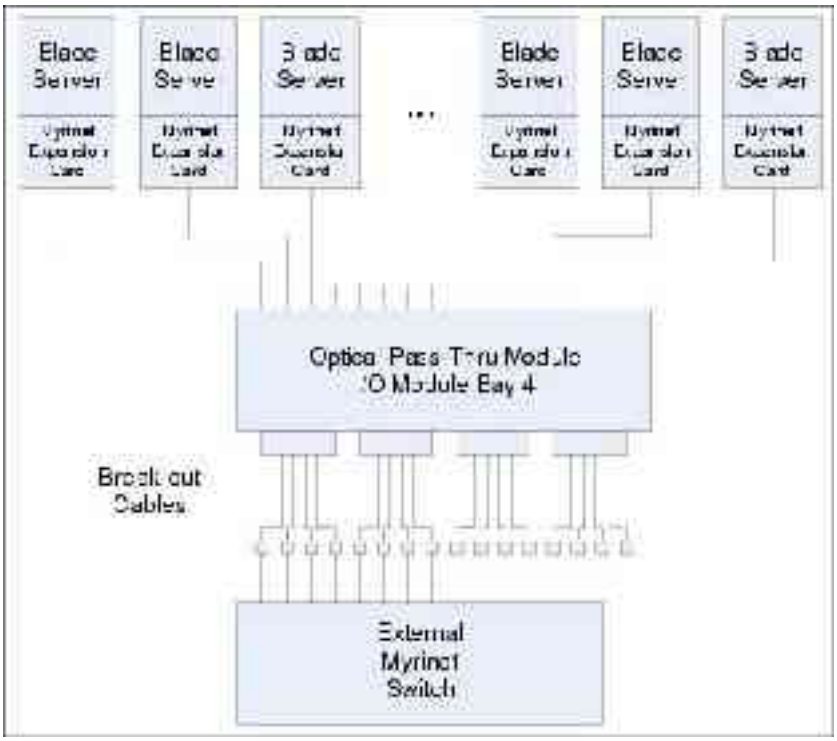
- Una agresiva predicción de saltos
- Ejecución en orden de hasta cinco operaciones por ciclo
- Ejecución fuera de orden de hasta diez operaciones por ciclo
- Renombrado de registros

El número total de instrucciones siendo tratadas en el nucleo del procesador en cualquier momento en el tiempo puede ser tan alto como 215.

Red de Interconexión

Las aplicaciones con memoria distribuida requieren el uso de una red de alta performance y baja latencia. Myrinet es una tecnología de red que cumple con estos requisitos y es la alternativa elegida para MareNostrum.

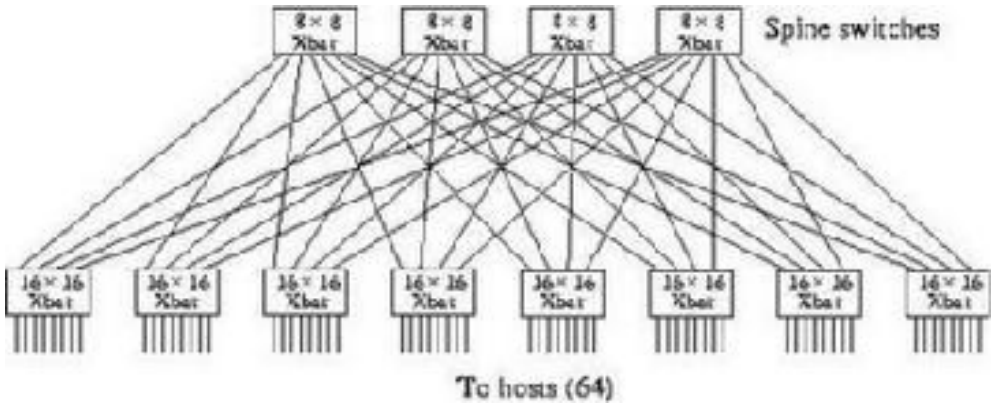
Se instaló una placa de red Myrinet en cada JS20 y un módulo "Optical Pass-Thru I/O" en cada una de las bahías de cada chasis. Luego se conectaron estos módulos a switches Myrinet externos para completar la topología como se muestra en la siguiente figura:



La infraestructura Myrinet puede ser utilizada para soportar herramientas de programación paralela tales como MPI o, en general, cualquier aplicación basada en protocolos IP ya que es posible asignar direcciones IP a una interface Myrinet. Por ejemplo, se puede soportar sistemas de archivos distribuidos como el "General Parallel File System (GPFS)" de IBM.

Myrinet es actualmente una de las alternativas a ethernet más utilizadas en redes de alta velocidad para clusters. Además del gran ancho de banda (cerca de 100 MB/s), la mayor ventaja se encuentra en que está implementada totalmente en espacio de usuario evitando de esta forma cualquier interferencia con el sistema operativo. Las últimas implementaciones de Myrinet sólo utilizan fibra óptica como medio físico, lo que brinda mayor velocida y menor tasa de errores.

Se ofrecen modelos de switches que tienen desde 8 hasta 256 puertos. Los switches de 8 y 16 puertos son full crossbars. Es posible construir una red de gran tamaño a partir de switches de 8 y 16 puertos y una topología Clos, como se observa en el siguiente ejemplo de 64 nodos:



La red de MareNostrum

Cuatro bastidores de switches Myrinet, que incluye 10 switches "CLOS 256+256" y 2 switches "Spine 1280" y el cableado Myrinet, todo ello en un sistema compacto, permiten que el proceso en paralelo sea más rápido, con una necesidad menor de hardware de conmutación. La fuente de alimentación rendundante y de intercambio dinámico garantiza una mayor disponibilidad. El sistema completo consta de 12 chasis con 2560 puertos. La uniformidad de estos puertos simplifica el modelo de programación, de tal modo que los investigadores podrán centrarse en sus proyectos y no en la arquitectura de interconexión.

Un switch CLOS 256+256 interconecta hasta 256 blades y se conectan con un switch Spine 1280 a través de 64 puertos. El Spine 1280, a su vez puede conectarse con hasta 10 switches CLOS a la vez.

Además, MareNostrum cuenta con una red gigabit ethernet utilizada para tareas administrativas y para la conexión con el mundo exterior.

Software:

El kernel linux 2.6 en su versión para la arquitectura PowerPC ofrece la característica de virtualización que permiten un particionamiento más flexible, un mejor balance de caraga y mayor escalabilidad. Además se aprovecha una utilidad llamada "Diskless Image Management (DIM)" que permite que los blades servidores obtengan todos los archivos de la distribución Linux desde los servidores de almacenamiento, utilizando la red del cluster. MareNostrum utiliza el pasaje de mensajes como principal modelo de programación, implementado por la librería MPI.

4. Conclusiones

Una comparación de las máquinas estudiadas con respecto al hardware disponible hace sólo algunos años permite observar una rápida evolución en varios aspectos de la computación paralela. Tanto la mejora de los medios de interconexión y el hardware de red como las prestaciones que ofrecen los nuevos procesadores superescalares de 64 bits hacen que los diseñadores y desarrolladores de software deban buscar contínuamente nuevas formas de explotar el poder de cómputo que implican estas máquinas. En función de eso, el software de código abierto, creado y mantenido por grandes comunidades, con la seguridad y flexibilidad que lo caracterizan y principalmente con la posibilidad de ser modificado, adaptado y mejorado por cualquier persona que lo deseé, se ha convertido en la opción elegida por la mayoría de los responsables de estas grandes máquinas. Como se comentó al principio, el kernel Linux en su versión 2.6, junto con la librería de pasaje de mensajes MPI, son la solución de código abierto que aparece en la mayoría de las supercomputadoras listadas en el ranking de www.top500.org.

Entre los modelos de programación paralela existentes, el pasaje de mensajes es el más utilizado. Y esto se debe, probablemente a que es el modelo que mejor se adapta a estas máquinas con miles de procesadores. Sin embargo existen soluciones que implementan memoria compartida distribuida entre una gran cantidad de procesadores, como es el caso de la Altix, que ofrece memoria compartida para 512 procesadores utilizando protocolos específicos implementados en hardware y modificaciones al kernel Linux.

5. Bibliografía

- Organización y Arquitectura de Computadores - William Stallings
- Advanced Computer Architecture. Parallelism, Scalability, Programability - Kai Hwang
- www.top500.org
- An Overview of the BlueGene/L Supercomputer - The BlueGene/L Team
- http://www.nas.nasa.gov/Users/Documentation/Altix/hardware.html
- http://sc.tamu.edu/help/altix/architecture.shtml
- http://www.embedded-computing.com/articles/woodacre/
- http://www-03.ibm.com/chips/products/powerpc/newsletter/jun2004/lead.html
- http://www.redbooks.ibm.com/redbooks/SG246342/wwhelp/wwhimpl/js/html/wwhelp.htm